REMBRANDT

Empowering Translational Research…

REpository of Molecular BRAin Neoplasia DaTa

Agenda

- Translational Research – Why do we care?
- GMDI – How we got here?
- Conceptual Model
- Gene Expression Use Case analysis
- Gene Expression Data analysis
- Wire Frames
- System Architecture
- Object Model
- Data warehouse design

Translational Research – Why do we care?

- Iressa Drug Case Study (at Harvard Medical School)
  - Targeted towards lung cancer
  - Phase II trial – A minority of patients showed dramatic tumor shrinkage
  - Phase III randomized trial – No survival improvement.
  - Patients with mutations in Iressa's target, EGFR, showed response to the drug.
  - Pharmacogenomics future is based on translational research

Reference: *Clinical Pharmacogenomics: Almost a reality; Modern Drug Discovery, August 2004*

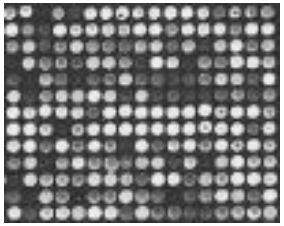
Scientific goals of GMDI

- Develop a molecular classification schema that is both clinically and biologically meaningful, based on gene expression and genomic data from tumors (Gliomas) of patients who will be prospectively followed through natural history and treatment phase of their illness

Secondary objective of GMDI

- Explore gene expression profiles to identify patient responsiveness
- Correlate gene expression profiles with discrete chromosomal abnormalities
- Identify a group of germ-line SNPs within immuno-regulatory genes that differentiate clinical outcomes
- Accurately assess the cellular biological activity based on protein (those involved in dynamic cellular activity, such as enzymes) content and activity

Rembrandt Knowledgebase

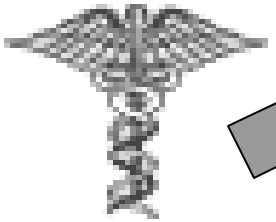


Expression array data

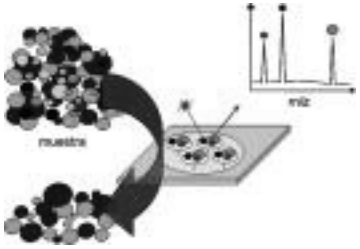
SNPArray data

Better understanding

Better treatments



Clinical data

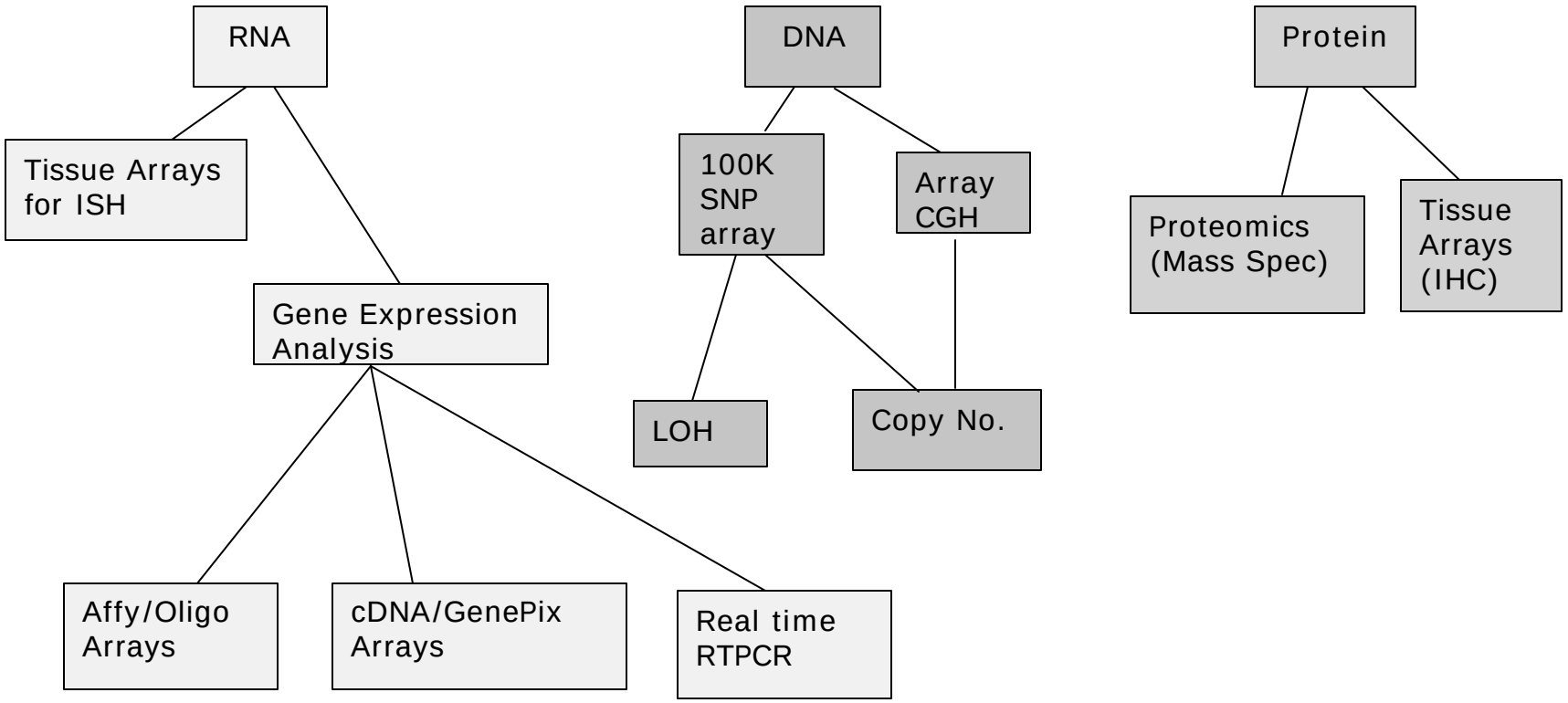


Proteomics data

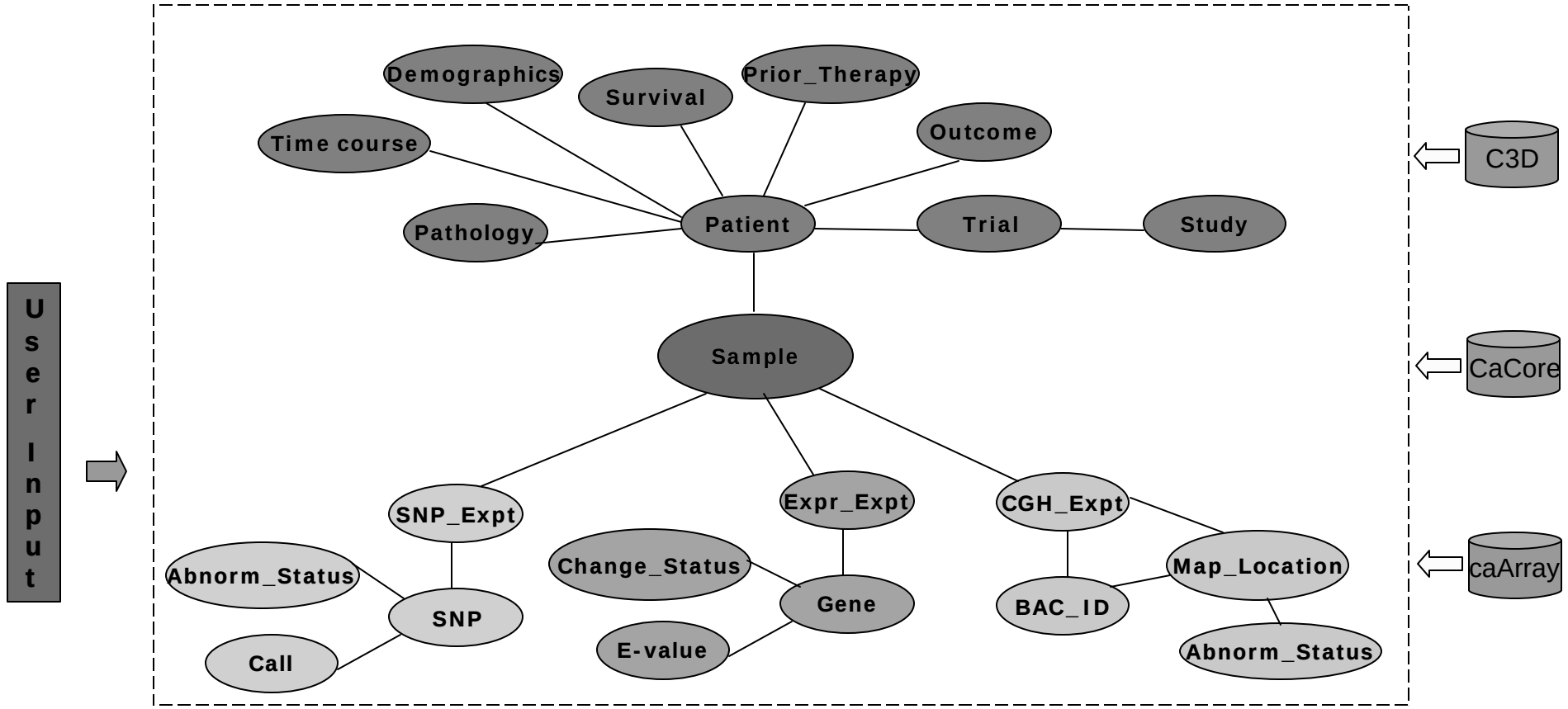
REMBRANDT Project Goals

- Produce a national molecular/genetic/clinical database of several thousand primary brain tumors that is fully open and accessible to all investigators (including intramural and extramural)
- Provide informatics support to molecularly characterize a large number of adult and pediatric primary brain tumors and to correlate those data with extensive retrospective and prospective clinical data

Functional genomics data in the knowledge-base



Conceptual Model



REMBRANDT will Leverage NCICB and caBIG Infrastructure Components

- Aligns with caBIG principles:
  - Open source
  - Open access
  - Syntactic and Semantic interoperability
  - Federated data
- NCICB Infrastructure
  - caARRAY gene expression data repositories and analysis tools
  - Cancer Genome Anatomy Project (CGAP) genomic tools
  - C3D Clinical Informatics System
  - caCORE Infrastructure (caBIO, EVS, caDSR)
- caBIG Infrastructure being delivered by caBIG workspaces

Typical Rembrandt Search

- Show me the tumors (Tumor samples) that have amplification and over-expression of Genes EGFR & Cyclin D1.
- Restrict the search to cases with
  - amplification confirmed by SNP Chip and CGH,
  - and over-expression confirmed by Oligo and cDNA Arrays
- Presentation of Results
  - Which genes are under-expressed respect to normal?
  - Do this subset of tumors have a better survival?
  - Do they segregate to a certain age group, geographical area or ethnicity?

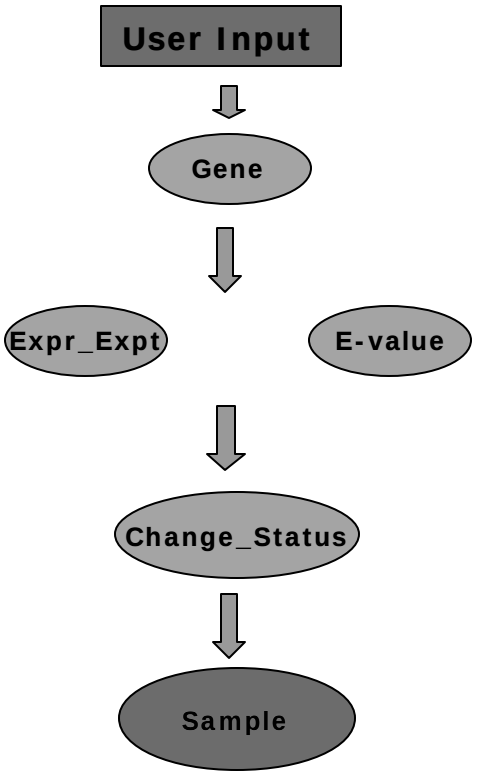
Inputs

- Collect Inputs to base search on:
  - Gene Names: EGFR & Cyclin D1
  - Fold Change threshold

    (to determine UP/DOWN regulation)

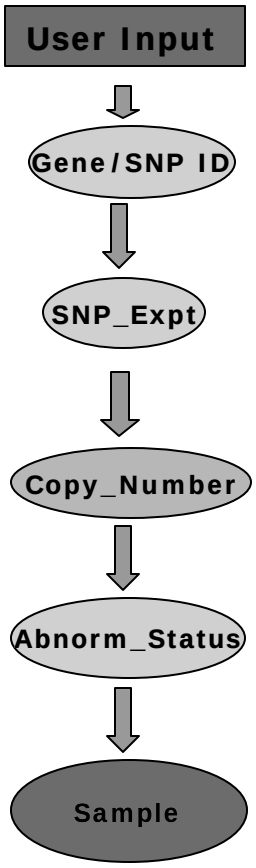
Microarray lookup (Affymetrics Platform)

1. Get reporter Ids for Gene EGFR & Cyclin (from caBIO)
2. Find all hybridization experiments associated with these reporter Ids.
3. Get Expression Values (E-values) for EGFR & Cyclin D1 in each hybridization.
4. Map E-Values to UP/DOWN Regulation status
5. Get samples

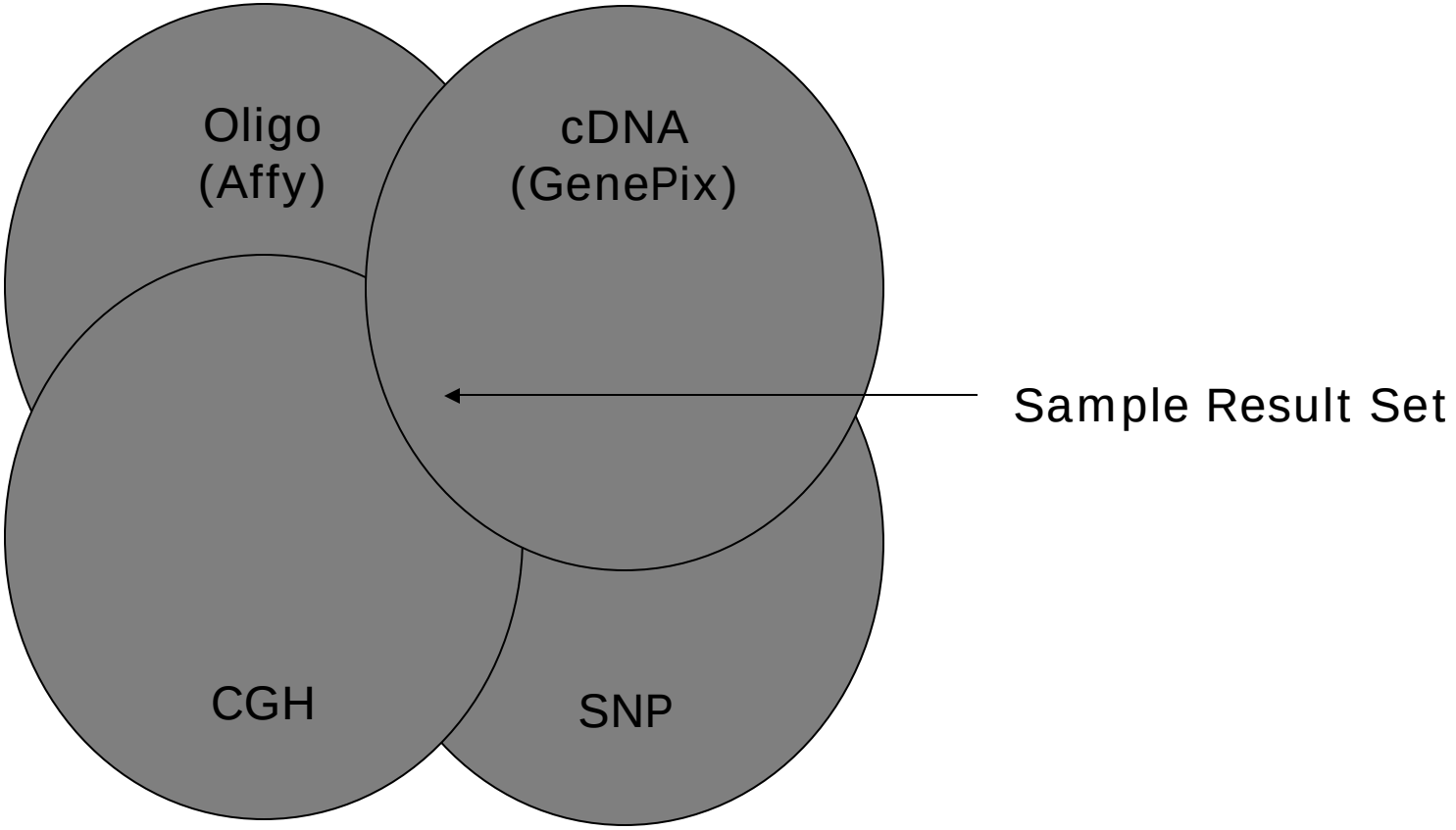


SNPArray lookup

1. Get SNP reporter Ids for Gene EGFR & Cyclin D1 (from caBIO)
2. Find all hybridization experiments associated with these reporter Ids.
3. Get Copy number for EGFR & Cyclin D from each hybridization.
4. Map Copy number to chromosomal abnormality
5. Get samples

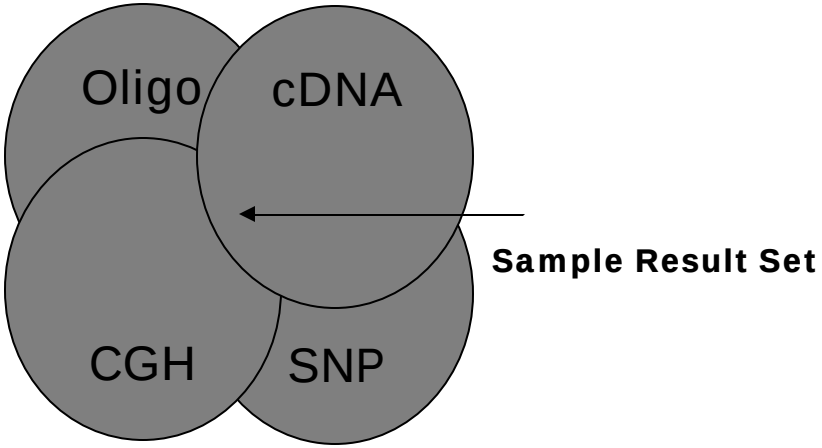


Search Results



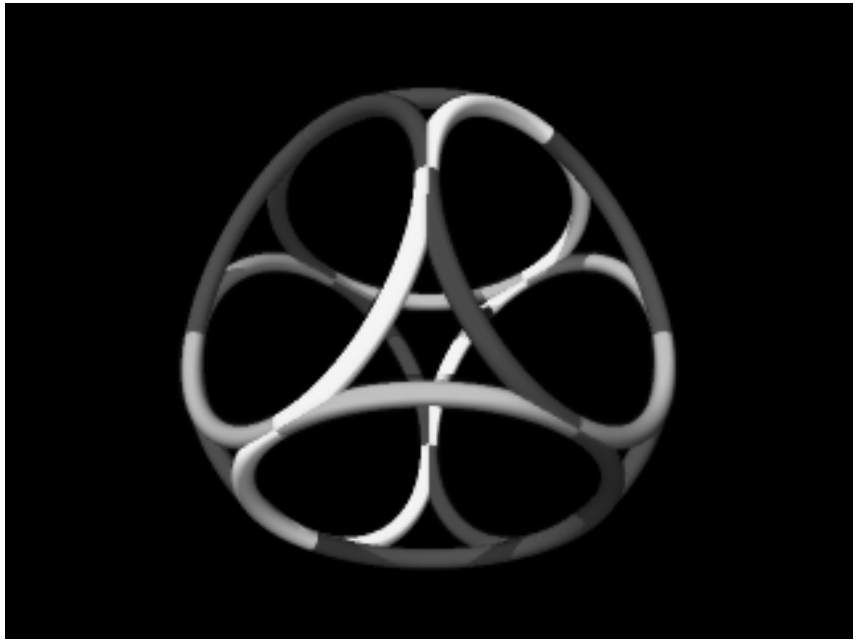
Search Result Presentation

- For each sample in the result set (shaded area) get the following:
  - all DOWN Regulated Genes
  - ✓ Survival Ranges
  - ✓ Age Group
  - ✓ Geographic Area
  - ✓ Ethnicity
- But how do you get all DOWN Regulated Genes?

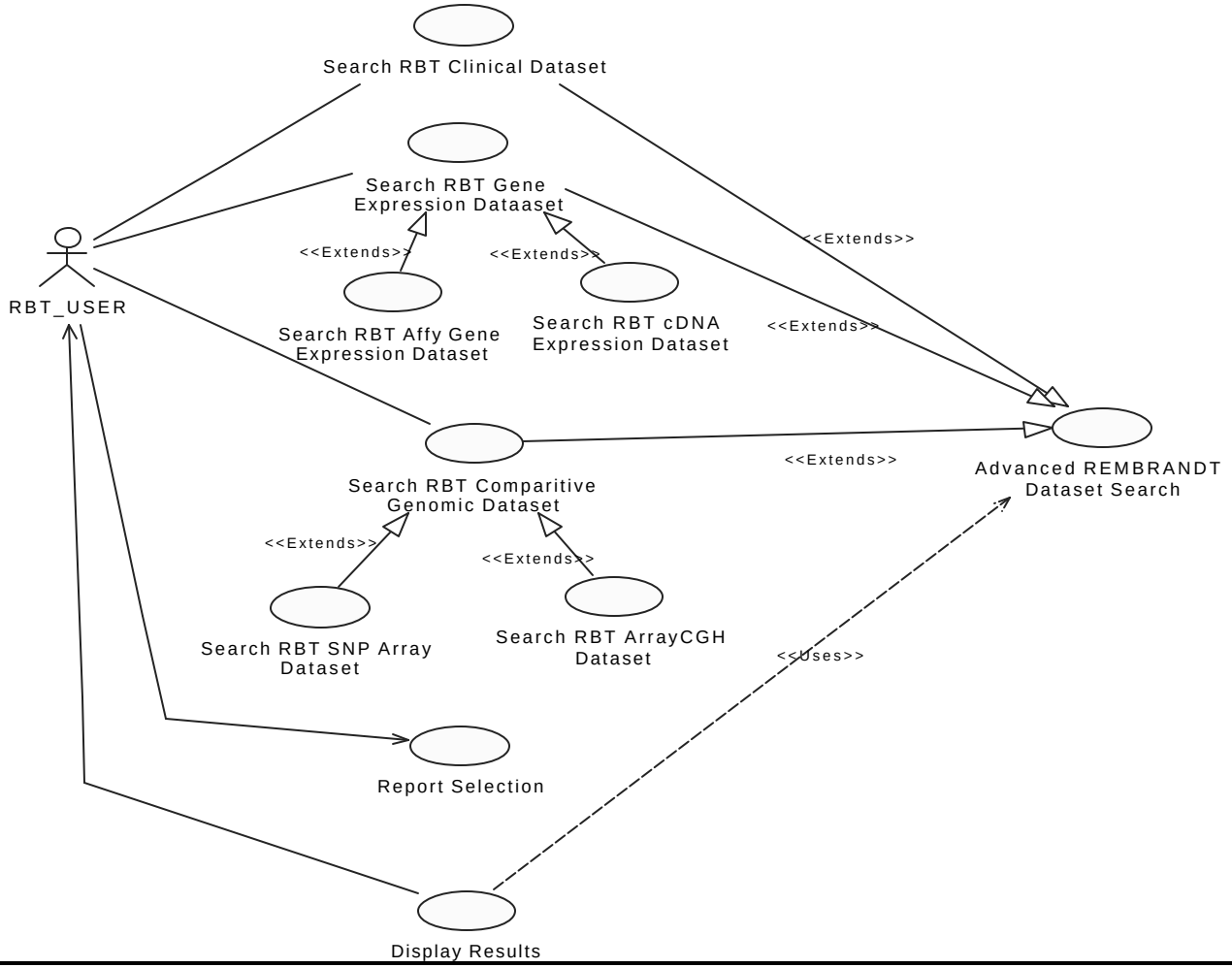


True Measure of Translation Research

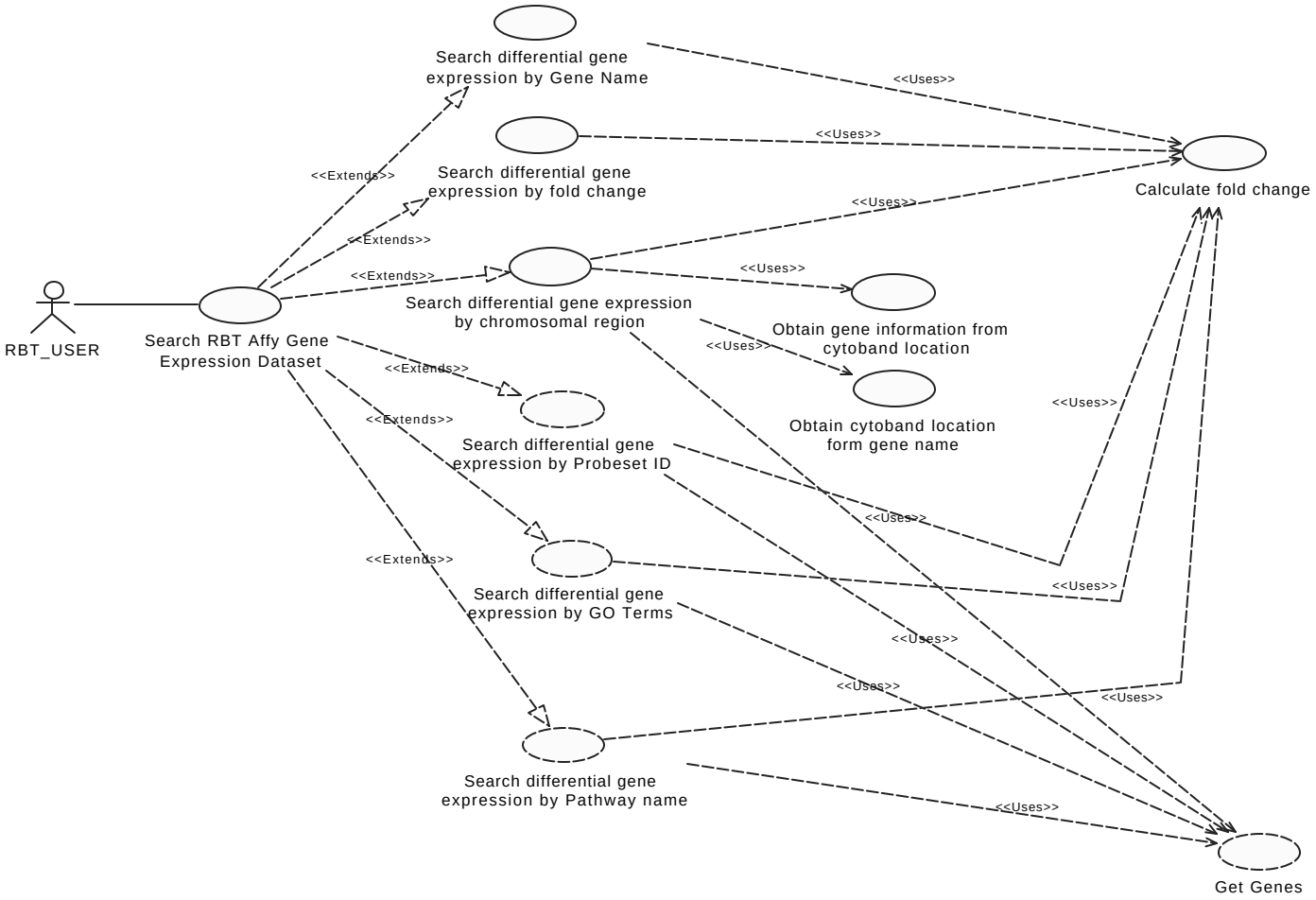
- To present the all DOWN Regulated Genes within each sample in the result set, we have to pivot the result set on its Gene Expression axis.
- All Translational Queries should allow the ability to easily pivot between:
  - Disease View
  - Patient / Sample View
  - Experiment/ Annotations View
  - Time Course View



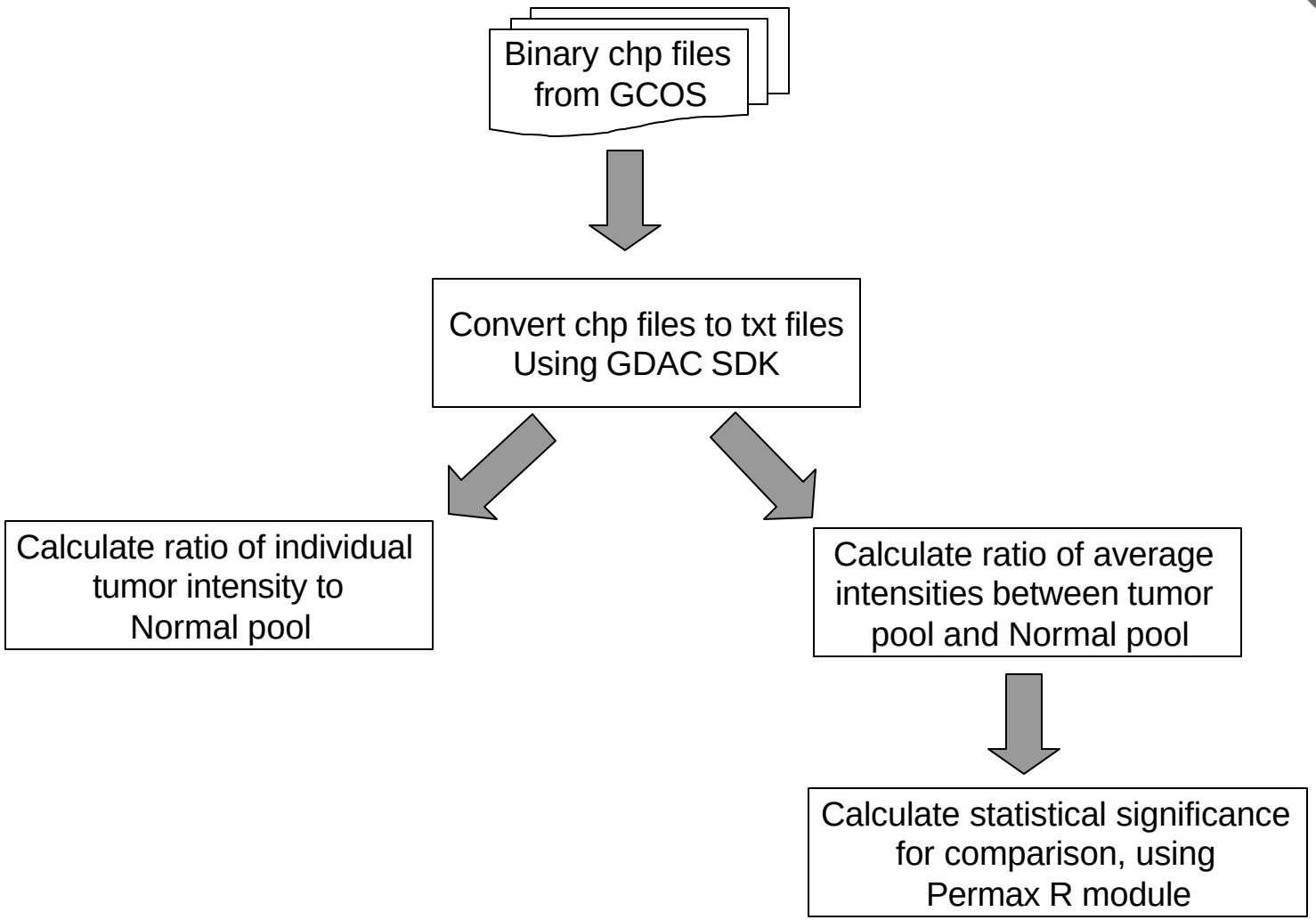
High-level Search Use cases



Gene Expression Search Use cases



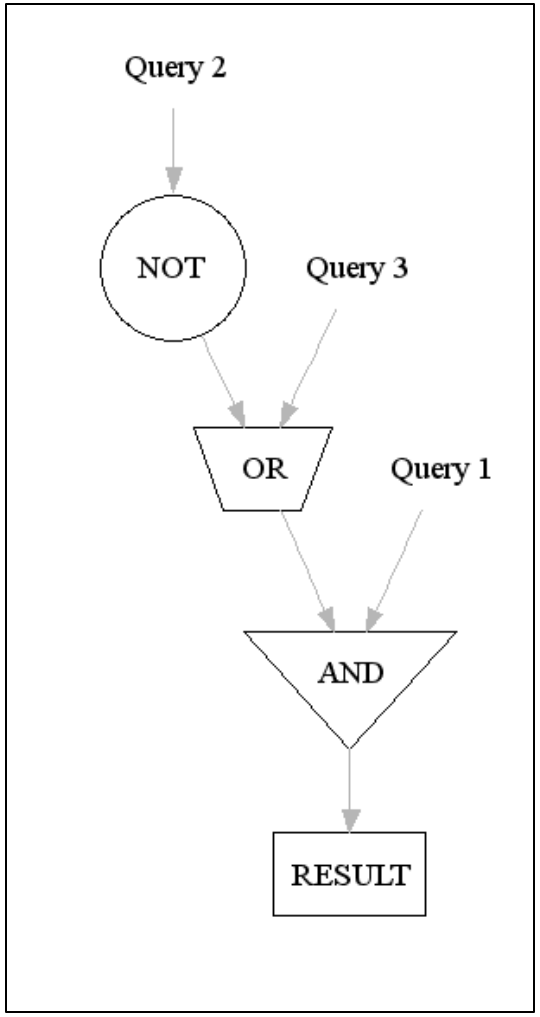
Gene Expression data analysis

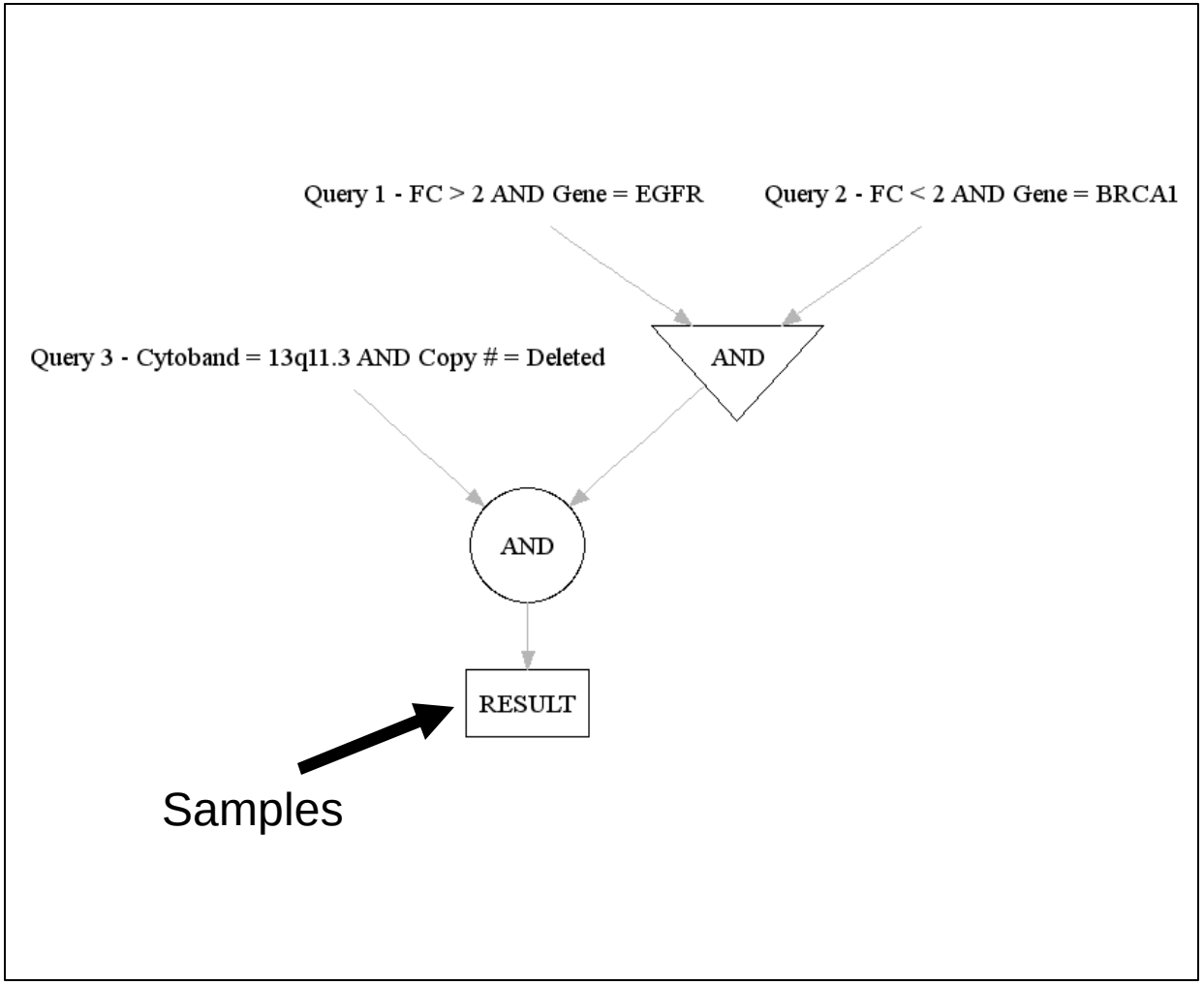


UI Wire Frames

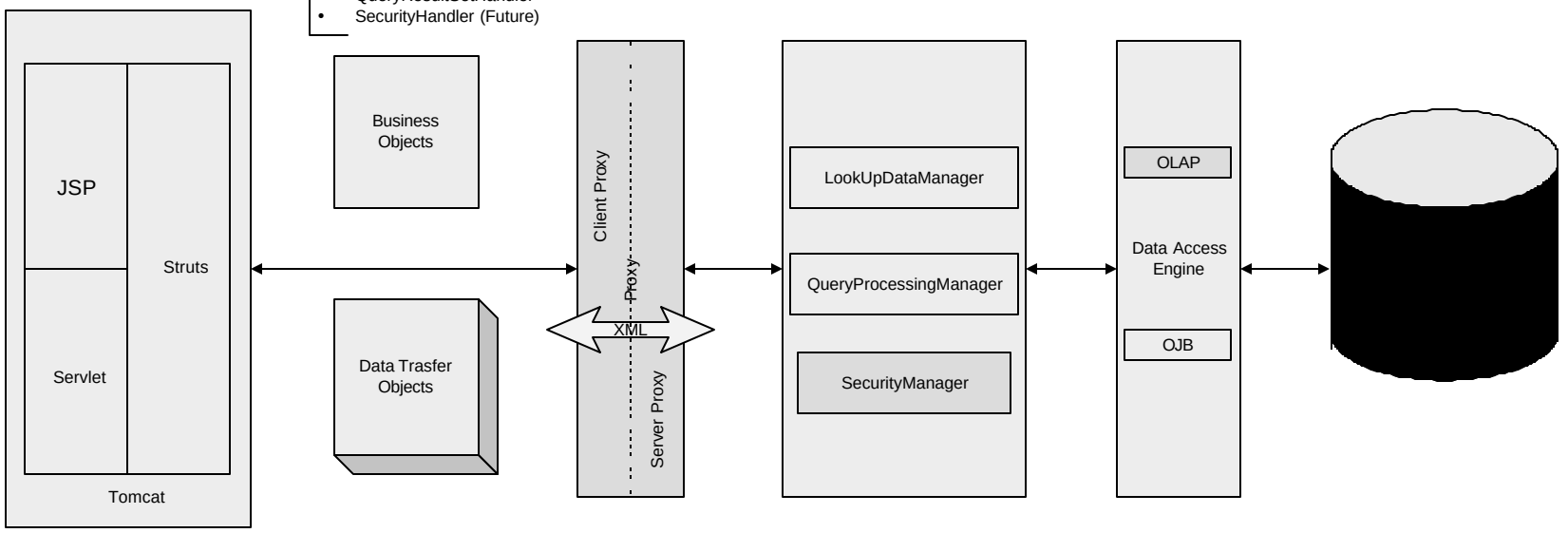
Link to Rembrandt search pages

Graphical representation of Queries



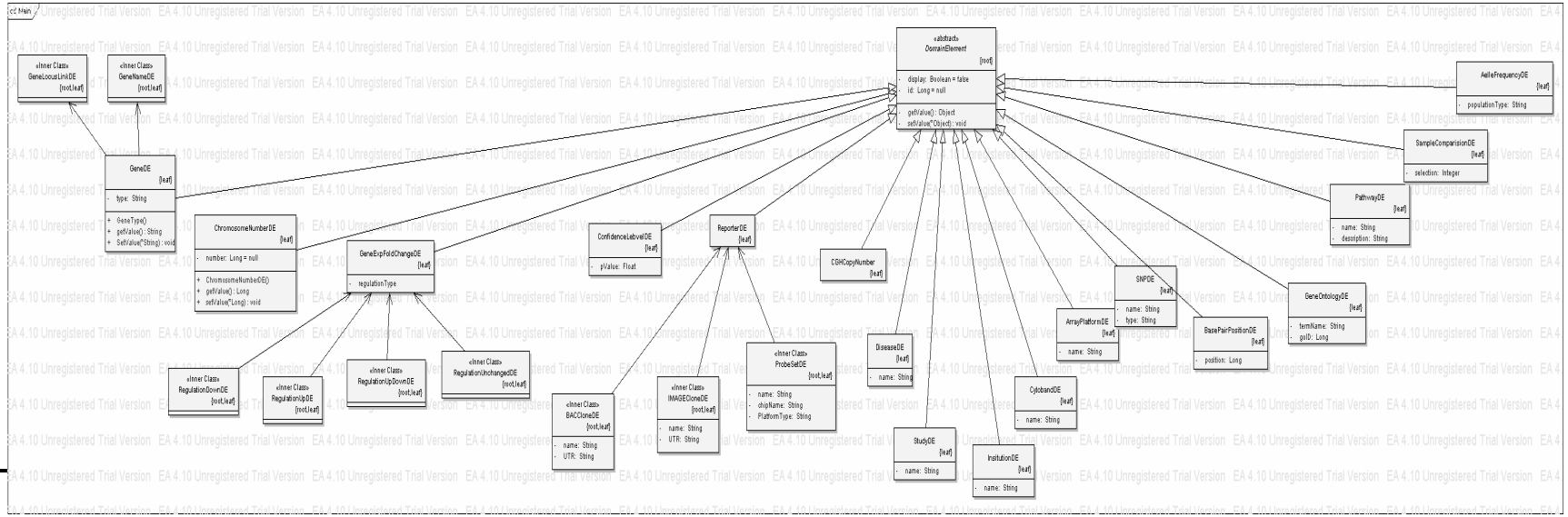


Architecture



Object Model

- DomainElement
  - Represents the basic elements involved in translational research space.
  - All queries, views and presentation objects are composed of domain elements
  - Provides strong type checking and validations



Database Schema

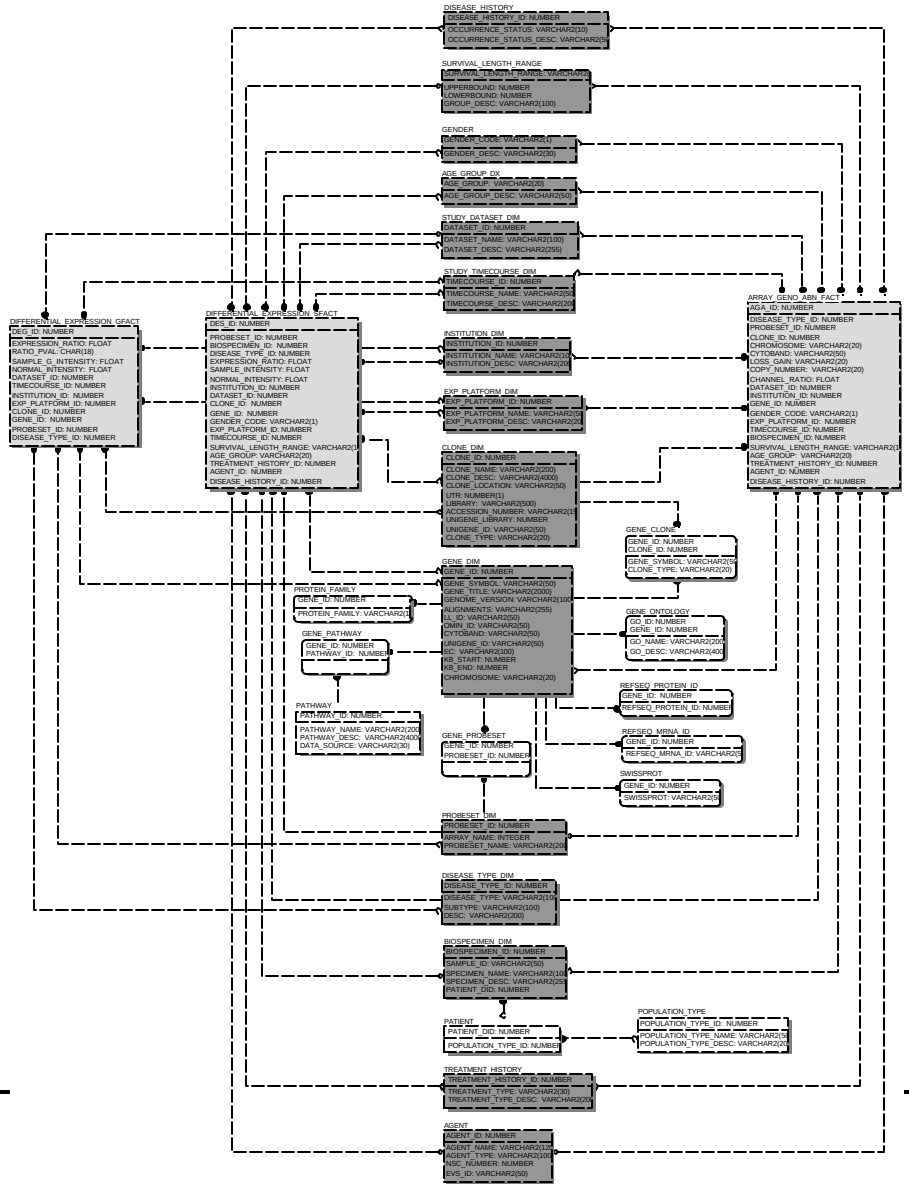
- Star schema
  - Is a generic, query optimized schema
  - A star schema consists of Fact tables and dimensions
  - Provides a highly de-normalized view of the data
  - Provides a data neutral framework from which queries can be executed with very fast results
  - Nautilus usage will help us validate our approach

Nautilus Database Schema

- Fact Table
  - Contains key performance indicators
  - Helps eliminate expensive joins from queries
  - In the future, if multi-dimensional measures are required, then our schema is extensible to allow us to perform OLAP queries

- Dimension
  - Dimensions are the categories of data analysis
  - When a report is requested "by" something, that something is usually a dimension.
    - For example, in a gene expression query, the two dimensions needed are genes (GENE_DM) and samples (BIOSPECIMEN _DM)

Database Schema



- I am done☺
- Questions☹